Building hierarchies and networks from MRDs of Tamil

S.Rajendran
Tamil University
Thanjavur 613 005

Introduction

There is rising interest in computational analysis of MRDs as a lexical resource for various purposes in the recent years (Wilsk et al, 1996:161-181). Existing dictionaries are obvious place to look for extensive linguistic information on lexical semantics. Exploiting MRDs for semantic analysis is a clear-cut effort, which aims at harvesting the work-years of effort by professional lexicographers. It is a well-known fact that a wealth of implicit information lies within dictionaries. This lexical information can be made explicit for use by programs. In dictionaries the lexicographers define words in terms of other words, which themselves have definitions elsewhere in the dictionary. The computational linguists have to "unpack" the meanings that have between encoded in dictionaries by lexicographers by converting them into a computationally tractable form. The motivation for creating lexical database in this way is to take advantage of the extensive lexicographic work that goes into creating the dictionary in the first place.

It is necessary to build a knowledge base that includes hierarchical information in order to process language effectively. It is not difficult to argue that the knowledge base should 'know' facts like a poodle IS-A dog, a dog IS-A mammal, and a mammal IS-A animal. In the knowledge base, poodles will have all the properties of dogs, and dogs will have all the properties of mammals, etc. and although there are differing opinions about whether this knowledge is inferred at that time of processing or inferred earlier and stored in the knowledge base, this is nonetheless crucial information which must be available for language processing. In general, the network/hierarchy efforts are more narrowly devoted to detailing hypernymic and hyponymic relations (at other times referred to as IS-A relation or AKO [a-kind-of] relations) between words and word senses defined in the dictionary. These are meaningful (semantic) relations explicitly intended to support inheritance. Dictionaries provide a rich source from which we can extract this kind of information automatically on a large scale.

The notion of automatically converting information available in MRDs into a "machine-tractable" one (i.e. Machine Tractable Dictionary, shortly MTD) is the central theme of this paper. The present proposal become relevant in the context of building a WordNet for Tamil in which the authors are presently involved. Though there are a few MRDs are available for Tamil, the most suitable among them for our work is *kiriyaavin taRkaalat tamiz akaraa*ti (KTTA).

Extraction of semantic information form dictionaries

Much of the information needed for our purpose can be found in a dictionary. The problem, of course, is that none of this comes without great effort. Every dictionary is printed in a different format, and every format is designed to be both

information-bearing and pleasing to the eye. This combination makes it very difficult to remove or reformat the strings and symbols in the dictionary automatically, to decide, by program, what is intended by the lexicographer, what to keep, and what to leave out. Extraction of semantic information from dictionaries entails certain assumptions about the extent of knowledge in a dictionary, about where that knowledge is located, and how that knowledge can be extracted from the language of dictionary definitions. These assumptions are methodological assumptions because they underlie the decisions made in choosing one method for semantic analysis rather than another. These assumptions are about sufficiency, extricability, and bootstrapping (Wilks et al.1990, 1996:141-142).

Sufficiency addresses the issue of whether a dictionary is a strong enough knowledge base for Tamil, specially as regards linguistic knowledge and, above all, the knowledge of the real world needed for subsequent text analysis. Different positions have been taken toward sufficiency within computational lexicography. Some researchers believe that there is not enough knowledge in dictionaries in principle; in other words, that certain specific semantic information is not available anywhere in a dictionary and hence must be derived from another, outside, source. Other researchers believe that dictionaries may indeed contain sufficient knowledge, although that knowledge may be implicit, because that knowledge can be made explicit by using information from entries in other parts of a dictionary.

Extricability is concerned with whether it is possible to specify a set of computational procedures that operate on an MRD and extract, through their operation alone and without any human intervention, general and reliable semantic information on a large scale, and in a general format suitable for, though independent of, a range of subsequent NLP tasks.

Bootstrapping refers to a process of collecting the initial information that is required by a set of computational procedures that are able to extract semantic information from the sense definitions in an MRD. The initial information needed is commonly linguistic, normally syntactic and case information, which is used during the processing of dictionary sense definitions into an underlying representation form which semantic information is then extracted. Boost strapping methods can be internal or external. Internal bootstrapping methods obtain the initial information needed for their procedures from the dictionary itself and use procedures to extract that information. By contrast, external bootstrapping methods obtain the initial information for their procedures by some method other than the use of those procedures themselves. The initial information may be from some source external to the dictionary or may be in the dictionary but impossible to extract without the use of the very same information. There are differences of opinion in computational lexicology regarding extricability and bootstrapping.

The straight forward approach to creating an MTD is to segment the dictionary into a collection of records, one for each word sense, composed of a set of labeled constituent fields, where each filed is roughly equivalent to a single meaningful element in a sense definition. The procedure makes the contents of the dictionary explicit, and is the so-called lexical database approach for forming an MTD. The more difficult approach is to process the labeled constituent fields in some way, in

order to retrieve the implicit knowledge that lies buried there. This is the semantic approach to creating an MTD.

Net work of relations

In addition to research into constructing network of genus terms form MRDs, various proposals in the past have suggested that different specialized link types should be added to the net-work. There are approaches to forming semantic networks from dictionary text using hand-coding have also been tried. The Princeton group (Beckwith et al. 1989) is now in the process of creating a network of nouns, verbs, and adjectives in accordance with psycholinguistic guidelines. The nodes in their work are synonym sets and its links are of many types. The nodes of their network are synonymy sets and its links are of many types. Within category of nouns there are both hypernym, hyponym, meronym (is-part), and holonym (reverse of meronym) links. Within the verb category are troponym (manner-of) and entailment (implicatgion) links. The following tables depicts various sense relations proposed in Tamil wordnet (Rajendran, 2002). In the Tamil wordnet, sense relations are established manually by using dictionaries.

Table of lexical relations for Nouns

Relations	**Subtypes**	**Example**
Synonymy		*puttakam* 'book' to *nduul* 'book'
Hypernymy-Hyponymy		*vilangku* 'animal' to *paaluuTTi* 'mammal'
Hyponymy-Hypernymy		*pacu* 'cow' to *paaluuTTi* 'mammal'
Holonymy-Meronymy	Wholes to parts	*meecai* 'table' to *kaal* 'leg'
,,	Groups to members	*tuRai* 'department' to *peeraaciriyar* 'professor'
Meronymy-Holonymy	Parts to wholes	*cakkaram* 'wheel' to *vaNTi* 'cart'
,,	Members to groups	*paTaittlaivar* 'captain' to *paTai* 'army'
Opposites	Antonymic (gradable)	*ndallavan* 'good person' to *keTTavan* 'bad person'
,,	Complementary	*iravu* 'night' *to pakal* 'day'
,,	Privative (opposing features)	*ahRiNai* 'irrational' to *uyartiNai* 'rational'
,,	Equipollent (positive features)	*aaN* 'male' to *peN* 'female'
,,	Reciprocal Social roles	*maruttuvar*'doctor' to *ndooyaaLi* 'patient'
,,	Kinship Relations	*ammaa* 'mother' to *makaL* 'daughter'
,,	Temporal Relations	*munnar* 'before' to *pinnar* 'after'
,,	Orthogonal or perpendicular	*vaTakku* 'north' to *kizakku* 'east' and *meeRku* 'west'

,,	Antipodal Opposition	*vaTakku* 'north' to *teRku* 'south'
Multiple opposites	Serial	*onRu* 'one', *iraNTu* 'two', *muunRu* 'three', *ndaanku* 'four'
,,	Cycle	*njaayiRu* 'Sunday' to *tingkaL* 'Monday' .. to *cani* 'Saturday'
Lexical association	Collocation	*cingkam* 'lion' to *karji* 'roar'
,,	Morphological relations	*paTi 'study'* to *paTittavan* 'educated man'
Compatibility		*ndaay* 'dog' to *cellappiraaNi* 'pet'

Table of lexical relations for Verbs

Relations	**Definition/sub types**	**Example**
Synonymy	Replaceable events	*tuungku* 'sleep' → *uRangku* 'sleep'
Meronymy-Holonymy	Events to super-ordinate events	*paRa* 'fly' → *pirayaaNi* 'travel'
Troponymy	Events to their subtypes	*ndaTa* 'walk' → *ndoNTu* 'limp'
Entailment	Events to the events they entail	*kuRaTTaiviTu* 'snore' → *tuungku* 'sleep'
,,	Event to its cause	*uyar* 'rise' → *uyarttu* 'raise'
,,	Event to its presupposed event	*vel* 'succeed' → *muyal* 'try'
,,	Event to its implied event	*kol* 'murder' → *iRa* 'die'
Antonym	Opposites	*kuuTu* 'increase' → *kuRai* 'decrease'
,,	Conversensess	*vil* 'sell' → *vaangku* 'buy'
,,	Directional opposites	*puRappaTu* 'start' → *vandtuceer* 'reach'

The present proposal of extracting these information form MRD automatically is a challenge.

Extraction of synonymy relation

The dictionaries of past tried to give synonymous items to the lexical item concerned. The earlier types of dictionaries in Tamil try to give synonymous term of the least understood lexical item. Even KTTA gives synonymous item to certain lexical items. This helps in establishing synonymous relations between lexical items.

ceey 'child' : *kuzandtai*

Extraction of hyponymy-hypernymy relation

Dictionary definitions of nouns are normally written in such a way that one can identify for each headword, a "genus term", and these are related via an IS-A relation. Information following the genus term, the differentia, serves to differentiate

the headword from other headwords with the same genus. Let us look at the the meaning of the *katti* found in KTTA.

> *katti* 'knife' *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa veTTum karuvi* 'a cutting instrument of flat meal plate with sharp edges '

In the definition *karuvi* 'instrument' is the genus term of the headword *katti* and *paTTaiyaana ulookattakaTTin oorangkaLil kuurmai uTaiyataaka uLLa* 'flat meal plate with sharp edges' and *veTTum* 'that which can cut' yields differentia. The standard technique for creating a semantic network form noun definitions is to identify the IS-A relationships of this type. The standard technique for creating a semantic network form noun definitions is to identify the IS-A relationships of this type.

We can view the process of building a semantic hierarchy of IS-A links from nouns definitions as two folds: (1) to find the genus term (or terms) in the definitions and (2) to disambiguate them.

Here we are concerned with the research projects in this area, the MRD's available for finding the hierarchies and networks for Tamil lexical items, and the degree of human intervention required.

The analysis of KTTA reveals us that about 90% of the noun definitions in the dictionary have the genus term, which is the head of the first noun phase (NP), and for these cases the standard method for identifying the genus term is to parse (or partially parse) the definitions and extract the head of the first NP. All researchers in this area have applied this method. However, all dictionaries seem to have a set of noun definitions that do not conform to this rule. Constructing taxonomies from the genus term thus forces one to take a stand on how to treat a rather large class of noun definitions which are not as 'standard' as the definition given above for *katti* 'knife'. The characteristic property of many of these definitions in that the head of the first NP (the usual place to find a genus term) seems vacuous, and another easily identifiable noun in the definition gives information about the headword. The analysis of definitions found in KTTA reveals us the following structure:

> *kattirikkool* : *iraNTu cama ndiiLa ulookapp paTTaikaLai kuRukkil onRin meel onRaaka iNaittu avaRRin kuurmaiyaana uTpakutiyaal tuNi mutaliyavaRRai veTTap payanpaTuttum oru karuvi*
>
> Word by word EG: scissors: two equal long metal strips_ACC across_LOC one_GEN on one_ADV having joined their sharp inside_part_by cotton_cloth things_of_that_sort_ACC cut_INF be_useful_RP one instrument
>
> 'an instrument made up of two long metal strips joined cross-wise which is used to cut materials like cotton clothes'

As for as Tamil is concerned the head occurs at the end of a clause. Here *karuvi* can be considered as the genus of the term *kattirikkool* 'scissors'. These items which

comes preceding the genus term can be considered as attributes. *payanpaTum* ‘that which is useful’ identifies the telic features of the lexical item, i.e. *veTTa* ‘for cutting’.

> *vaazai* 'plantain tree’ *inippuccuvai mikundta pazattait taruvatum uritteTukkak kuuTiya maTTaikaLaalandtum periya ilaikaLai uTaiyatu maana our vakai maram*
>
> EG: ‘...a kind of tree’

The key word *oru vakai* ‘a kind of’ helps in finding the genus term *maram* ‘tree’.

Some genus terms are synonymous in nature. The following example will illustrate the point.

> kuLir paanam: kuLira vaikkappaTTa paana vakai
> EG: ‘the drink which is caused to become cool’

In this definition the word *paanam* ‘drinks’ which comes before the key word *vakai* ‘kind of’ is the genus term.

Some times the genus term is far removed form the concerned word conceptually.

> *kaTivaaLam* ‘bridle with bit and reigns’: (*kutiraiyin ooTTattaik kaTTupptuttum vakaiyil) vaayilum talaiyilum poruttappaTum ndiiNTa vaaruTan kuuTiya caatanm*
>
> EG: ‘a thing which is a long strap joined to a horse’s mouth and head in order to control its running movement’

In this kind of description, the genus term identified *caatanam* ‘thing’ is far removed from the immediate genus.

Establishing meronymy-holonymy relation

Part-and-whole relation (i.e. meronymy-holonymy relation) in the following case is difficult to identify, though the genus term can be easily identified.

> viral ‘finger’: *kaiyin allatu kaalin mun pakutiyil aindtaakap pirindtuaLLa, ndakattooTu kuuTiya ciRiya uRuppu*

The genus term *uRuppu* is explicit. But its meronymic relation is with *kai* ‘hand’ or *kaal* ‘leg’ is difficult to establish though the genitive *–in* ‘of’ can help us to infer that the holonym of *viral* ‘finger’ is *kai* ‘hand’ or *kaal* ‘leg’. *pakuti* ‘part’ is the key term which can help us to establish meronymy-holonymy relation.

> *kaN* ‘eye’ : *paarppataRkup payanpaTum uRuppu*
> Word by word EG: seeing_for be useful_RP body_part
> ‘the body part useful for seeing’

In this definition *uRuppu* 'body part' is the genus of *kaN*. It more or less follow the previous pattern. The telic feature is given and other attributive or contentive feature is not given. *kaN* is the polysemous word and the identification of telic feature diambiguate it form other selecting other genus.

Made of relation

Some times the 'is made up of' relation can be established between lexical items by means of dictionary definition. The key word could be *ceyyappaTTa* 'that which is made'

> *kaLLippeTTi: caatikkaay marappalakaikaLaal ceyyappaTTa oru vakaip peTTi*
> EG: a kind of box made up of deal wood'

The genus term is *peTTi* 'box'. *peTTi* is in 'is made up of 'relation with *caatikkaay marppalakaikaL* 'deal wooden planks'.

Conclusion

The identification of a satisfactory genus term is not straightforward in all cases, and the problem associated with classifying the relationships expressed are difficult and numerous. A through study of this shadowy area is necessary in order to make optimal use of the semantic information available in MRDs.

Reference

Busa, F., Calzolari, N. and Lenci, A. 1998. "Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP." In

Cruse, D.A. 1986. Lexical Semantics. Cambridge, England: Cambridge University Press.

------. 2000. Meaning in Language: An Introduction to Semantics and Pragmatics. Oxford New York: Oxford University Press.

Hearst, M.A. 1999. "Automated Discovery of WordNet Relations." In Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. Cambridge, Mass. MIT Press.

Pustejovky, J. 1996. The Generative Lexicon. Cambridge, Mass. MIT Press.

Pustejovky, J. and Boguraev, B. 1993. "Lexical Knowledge Representation and Natural Language Processing." In F. Pereira and B. Grosz (eds.), Natural Language Processing. Cambridge, Mass.: MIT Press.

Rajendran, S. "Preliminaries to the preparation of a word net for Tamil"

Ruimy, N., Gola, E., and Monachini, M. "Lexicography Informs Lexical Semantics: The SIMPLE Experience." In

Wierzbica, A., 1980. Lingua Mentalis: The Semantics of Natural Language. New York: Academic Press.

Wilks, Y.A., Fass. D.C., Guo, C.M. McDonald, J.E., Plate, T. and Slator, B.M. 1990. "Providing machine tractable dictionary tools." In Pustejovksy, J. (ed.) Semantics and the Lexicon. Cambridge, Mass.: MIT Press.

Wilks, Y.A., Slator, B.M. and Guthrie, L.M. 1996. Electric Words: Dictionaries, Computers, and Meanings. Cambridge, Mass.: MIT Press.